23-5-2011

TIN-a combinatorial compound collection of synthetically feasible multicomponent synthesis products.

Kristl V. Dorschner
Royal College of Surgeons in Ireland

David Toomey
Royal College of Surgeons in Ireland

Marian P. Brennan
Royal College of Surgeons in Ireland, mbrennan3@rcsi.ie

Tim Heinemann
Royal College of Surgeons in Ireland

Fergal J. Duffy
University College Dublin

See next page for additional authors

Authors
Kristl V. Dorschner, David Toomey, Marian P. Brennan, Tim Heinemann, Fergal J. Duffy, Kevin B. Nolan, Dermot Cox, Mauro FA Adamo, and Anthony J. Chubb

TIN – a combinatorial compound collection of synthetically feasible multicomponent synthesis products.

Kristl V. Dorschner,[†] David Toomey,[†] Marian P. Brennan,[†] Tim Heinemann,[†] Fergal J. Duffy,[Ψ] Kevin B. Nolan,[‡] Dermot Cox,[†] Mauro F. A. Adamo,[‡] Anthony J. Chubb[†Ψ].*

University College Dublin, Clonskeagh, Dublin 4, Ireland

anthony.chubb@ucd.ie

TIN Is Not commercial.

* Corresponding author phone: +353 (0) 1 7165390; email: anthony.chubb@ucd.ie.

† Molecular and Cellular Therapeutics Department, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland.

‡ Centre for Synthesis and Chemical Biology, Pharmaceutical and Medicinal Chemistry Department, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland.

Ψ School of Medicine and Medical Sciences, University College Dublin, Complex and Adaptive Systems Laboratory, 8 Belfield Office Park, Clonskeagh, Dublin 4, Ireland

ABSTRACT

The synthetic feasibility of any compound library used for virtual screening is critical to the drug discovery process. TIN, a recursive acronym for 'TIN Is Not commercial', is a virtual combinatorial database enumeration of diversity-orientated multicomponent syntheses (MCR). Using a 'one-pot' synthetic technique, 12 unique small molecule scaffolds were developed, predominantly styrylisoxazoles and bis-acetylenic ketones, with extensive derivatization potential. Importantly, the scaffolds were accessible in a single operation from commercially available sources containing R-groups which were then linked combinatorially. This resulted in a combinatorial database of over 28 million product structures, each of which is synthetically feasible. These structures can be accessed through a free web-based 2D structure search engine, or download in SMILES, MOL2 and SDF formats. Subsets include a 10% diversity subset, a drug-like subset and a lead-like subset that are also freely available for download and virtual screening (http://mmg.rcsi.ie:8080/tin).

KEYWORDS

Chemoinformatics, combinatorial chemistry, multicomponent synthesis, diversity orientated synthesis, drug discovery, derivatization, database, structure search, free access.

BRIEFS

Enumeration of a virtual combinatorial compound library of 28m small molecules that can be synthesized using simple multicomponent methods.

INTRODUCTION

Virtual Screening (VS) is a viable and rapid method to search for ligands and antagonists of novel target proteins.[1-3] While the costly high throughput screening (HTS) of hundreds of thousands of synthesized compounds remains the predominant method used by the pharmaceutical industry, VS is steadily growing in importance and is increasingly being used to support drug discovery efforts.[4-6] Large databases of millions of compounds can be screened relatively rapidly using docking software, pharmacophore based substructure matching, or descriptor based filtering rules.[5] Ligand databases include collections of commercially available molecules such as the ZINC database compiled by Irwin and Shoichet,[7] or virtual representations of either 'in house' compound collections or commercially available compound libraries. Top scoring hits are then purchased and tested *in vitro* and may yield lead compounds effective against the target protein.

Development of the lead compound into a drug candidate may, however, be limited by synthetic complexity which reduces the scope for derivatization of these lead compounds. Derivative synthesis involving numerous steps, some or many of which may be patented, could result in poor yields and unpatentable drug candidates. This may necessitate the redesign of the synthetic route with concomitant delays in the drug discovery pipeline. Furthermore, the synthesis pathway may limit the available choice of derivatives due to incompatible side reactions. Synthetic feasibility of the compound library is thus critical to the drug discovery process.

A few years ago, Adamo's group initiated a program of research aimed at generating chemical diversity through multi-component reactions. This approach to the development of diversity-oriented syntheses was based on the generation of building blocks that contain multiple functionalities which could be selectively reacted. Therefore, it was hypothesised that a scaffold containing a number of functionalities m which could be *selectively* reacted through a number of n reactions would have generated diversity in $D = mn$ directions.[8] This study generated several one-pot multicomponent reactions, the synthetic details of which were described in previous papers.[8-13] Diversity is achieved by substituting each of three or four substituents at one time and keeping the synthetic strategy already

optimized constant. In this fashion synthesis of opportunely modified structures could be achieved within hours/days instead of weeks/months. While this is clearly a rich source of structural diversity, the technology remains in the domain of the skilled medicinal chemist and is not readily available for chemoinformatics exploration. We have thus enumerated a significant fraction of the possible derivatives that can be constructed using each of numerous multicomponent synthesis methods, resulting in over 28 million virtual compounds. It is important to note that each of these 28 million compounds is theoretically synthetically feasible, as examples of each scaffold structure have been synthesized, characterized, and described in the literature.[8-13] To the best of our knowledge, this is the first publicly available database of synthetically feasible combinatorial compounds that can be produced using the multicomponent 'one-pot' synthesis method (http://mmg.rcsi.ie:8080/tin). A number of subsets ('drug-like', 'lead-like' and 10% diversity) are available for download and incorporation into docking or virtual screening studies, while the remaining analogues are available through a web based search engine. Synthetically feasible analogues of lead compounds can thus be found using the 2D structural search. This will provide researchers with a set of readily synthesizable derivatives for lead optimization.

MATERIALS AND METHODS

Software used. The modelling program used for preparation of virtual molecular structures was MOE (Molecular Operating Environment) from the Chemical Computing Group, Montreal, Canada. The 2D search is performed using OpenBabel (www.openbabel.org). Statistics were prepared using a MySQL (www.mysql.com) database of the full library on an Ubuntu Linux server (www.ubuntu.com), and Toad (www.toadsoft.com) as the MS Windows interface. The website was programmed in Java (www.java.com), using NetBeans IDE (www.netbeans.org). Structural fingerprints for comparison to the ZINC database were generated using the Chemistry Development Kit java package (http://sourceforge.net/projects/cdk/).[14]

Synthesis of scaffolds MA01-MA21: The basic scaffold structure of each synthetic route is shown in Figure 1, while Schemes 1-12 provide more detail regarding the possible permutations of these

scaffolds. Compounds **MA01** (Scheme 1) were obtained by reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**), a β-diketone (**K**) in the presence and hydrazine or hydroxylamine (**X**) in the presence of piperidine followed by treatment with aqueous sodium hydroxide.[12] Compounds **MA02** (Scheme 2) were obtained by reacting equimolar amounts of 2,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**), a β-diketone (**K**) and hydrazine or hydroxylamine (**X**) in the presence of piperidine.[12] Compounds **MA03** (Scheme 3) were obtained by reacting **MA02** with $SnCl_2$ followed by treatment with an *in situ* generated acyl chloride. This synthesis is as yet unreported and will be described elsewhere. Compounds **MA04** (Scheme 4) were obtained by reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**) and a β-diketone (**K**) in the presence of substoichiometric amounts of piperidine base.[8] Compounds **MA05** (Scheme 5) were obtained by reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**) and a β-diketone (**K**) in the presence of two equivalents of piperidine base.[8] Compounds **MA06** (Scheme 6) were obtained by reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**) and a β-diketone (**K**) in the presence of substoichiometric amounts of piperidine base.[11] Compounds **MA07** (Scheme 7) were obtained in two steps reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**), and a β-diketone (**K**) in the presence of substoichiometric amounts of piperidine base, then adding DCC and amine. This synthesis is unreported and will be described elsewhere. Compounds **MA08** (Scheme 8) were obtained in two steps by reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**), and diethyl malonate (**K**) in the presence of substoichiometric amounts of piperidine base, then treating the resulting adduct with dilute HCl and excess $SnCl_2$.[11] Compounds **MA11** (Scheme 9) were obtained in two steps by reacting equimolar amounts of 3,5-dimethyl-4-nitroisoxazole (**Isi**), an aromatic aldehyde (**Ald**), and nitromethane (**K**) in the presence of substoichiometric amounts of piperidine base, then treating the resulting adduct with dilute HCl and excess $SnCl_2$[13]. Compounds **MA16** (Scheme 10) were obtained by reacting bis acetylenic ketones with

equimolar amounts of an opportune enamine.[9] Compounds **MA20** and **MA21** (Scheme 11 and 12) were prepared by reaction of bis acetylenic ketones and an opportunely substituted Boc-amidrazone in the presence of excess trifluoroacetic anhydride.[10]

Creating Constituent Databases. Creation and collation of initial constituent databases was performed manually to prevent common automation problems occurring early in the database creation process. Compounds which were identified as possible constituents in a scaffold-based molecule in the initial search were then searched for by product number in the electronic Sigma-Aldrich structure data file (SDF) catalogue. When a product was matched, constituent molecules were separated based on specific substructures, resulting in 11 databases (see Table 1). The Sigma SD files were planar and did not account for chirality. Hydrogen atoms were added where necessary and compounds allowed to relax using the steepest decent algorithm and the MMFF94x forcefield with default settings in MOE. Alternative, more efficient, options for 3D topology creation that are free for academia include the CORINA[15] and OMEGA[16] systems. To ensure correct stereochemistry we manually checked each molecule. If necessary, a molecule's chirality was changed to reflect that described in the full compound name. Where only racemic mixtures were available, or the full name could not be established, all combinations of chirality were manually stored in the database under the same name and catalogue number but with each set of chiral combination noted. All molecules with more than four chiral centers were discarded. Occasionally the structure needed to be reconstructed in MOE using the 2D drawing on the Sigma-Aldrich website as a guide. Where the product number differed from that on the website, this was updated in the constituent database to reflect the website nomenclature.

Creating Scaffold Databases. Numerous scaffolds were created on the basis of this synthetic technique with a high level of variety in their basic structure, as seen in Schemes 1-12. Various elements of the scaffold structure are variable while the basic structure generally remains consistent. This means that for every basic scaffold there are a number of different second generation scaffolds which can be created from it. This variety allows for even greater combinatorial expansion when combined with reagent (R-group) databases. Each scaffold possibility was constructed using MOE and then assigned a

name according to a specific nomenclature described in Schemes 1-12 and described in more detail here. Each scaffold has a basic structure which is indicated by the title **MA00**, where 00 represents a number which defines the synthetic route. **K** groups (depicted in blue in the Schemes) are represented by K00, and **X** groups (depicted in red or green) are represented by X00. For example, Scaffold **MA01** has fifteen different second generation scaffolds which are all defined by the combination of **MA**, **K**, and **X** groups present in the title. Therefore, MA01K02X03 indicates that this scaffold is made up of the basic parent scaffold of MA01, its variable R-groups which form a backbone are made up of $-CF_3$ groups, and the N-X group is in this case a methyl hydrazine ($N-N-CH_3$). If, as in Scaffold **MA02**, there was more than one **X** group present, the **X** groups were differentiated as being either **XA00** or **XB00**. Which group is considered A and B is noted in the Schemes. Individual databases of both X and K groups were also created to serve as a key marrying structure to name in MySQL backtracking of the component parts. Finally, each scaffold has an attachment site to which a subset of compounds with specific reactive groups will bind. For example, **MA02** molecules bind to aldehydes, as can be seen in Scheme 2. This was represented on the virtual scaffold as a 'port' with which an R-group molecule could later link. In some instances, for example **MA01**, some scaffolds have more than one 'port', often for different types of molecules. In this instance again, the 'ports' are differentiated as A1-An, where n represents an integer; these numbers must be labeled consecutively. Because of the large number of scaffold permutations associated with **MA03** (75 unique descendant scaffolds), the **MA03** database was split into 8 different databases called MA03(a)-(h). MA03(h) database contains only 5 scaffolds while the others contain ten. All scaffolds used can be downloaded from the TIN website (http://mmg.rcsi.ie:8080/tin).

Clipping R-Groups. Abbreviations used: aldehyde (**Ald**); alcohol (**Alch**), carboxylic acids (**Ac**), amines (**NHAmi**). Because each compound within the constituent database was screened initially by chemists with the one-pot synthesis in mind, each constituent was grouped to need the same 'clip' to prepare for linking with a scaffold. For example, scaffold **MA01** requires an aldehyde group which is seen in the resulting molecule as simply the Ald constituent as the carbonyl element of the aldehyde is removed in the chemical process. Therefore, when preparing the Ald databases of constituent molecules

to become a database of R-groups ready for combinatorial linkage, a 'clip' removed the carbonyl group and replaced it with a 'port' which is ready to dock with the scaffold resulting in the final product. Ports on the R-Groups are labeled with A0. Each 'clip' was designed with the synthetic process in mind so that the final virtual R-group, when linked, would be representative of the final product of chemical synthesis. These 'clips' are represented as SMILES strings in Table 1. Each clip was then assigned a unique name and number which allow for easy identification based on its chemical properties. For example, an aliphatic alcohol present in the Alch database would be named **Alch_naro_nnnn**, where n is an integer. This unique tag was linked to the R-group so that in subsequent molecular construction, it would always be present in the name of the final molecule, indicating which R-group was involved in making that product. The alcohol (**Alch**) subset was further divided into either phenols (**Alch-aro**) or aliphatic alcohols (**Alch-naro**), each of which were clipped to remove either a single hydrogen atom or the whole hydroxyl group. Only aromatic aldehydes were included in the aldehyde group (**Ald**). Carboxylic acids were subdivided into aromatic (**Ac-aro**) and aliphatic (**Ac-naro**) groups. Amine groups (**NHAmi**) were also divided into aromatic amines (**NHAmi_aromatic**) and the remainder (**NHAmi_naromatic**), both of which were further subdivided into either primary (**-NH2**) or secondary (**-NH**) amines.

Combinatorial Linking of R-groups to Scaffolds. By this point, both the scaffolds and the constituent R-groups have a 'port' with which they are able to link; combining to form a single virtual molecule. Combinatorial linking was done using the Quasar:Combigen function in MOE. For each linking process, a log file detailing any errors was generated. If possible, all errors were corrected manually, otherwise the compound was deleted.

Characterizing Output Databases. With construction complete, each output database was subjected to a variety of refinements. Each entry within each database was assigned a unique tag of **RCSI_nnnnnnnn**, where n is an integer. This tag was added to the original name of the molecule which consisted of the scaffold name and R-group identifying tags. Each database was then subjected to a steepest decent energy minimization process using the MMFF94x forcefield with default settings and

the addition of hydrogens where necessary. This was done using in parallel the SMP function and MOE batch mode on six Dell Dual Xeon Pentium 6 machines with 2Mb RAM, each operating Linux RedHat.

Molecular descriptors (listed in Supplementary Table S1) were then calculated for each output molecule. This descriptor calculation was done using functions in MOE's QuaSAR_Descriptor function in MOE/Batch. All the data are supplied in the download files. The descriptors were used to calculate the Lipinski 'drug-like'[17, 18] and Oprea 'lead-like' rules,[19] depicted with a binary '1' or '0'. In addition, substructure diversity selection was used to provide users with a smaller 10% diversity subset. All virtual molecules are available for download (in SMILES format only) and include the descriptor calculations.

The potential energy was calculated as a final check of the MOE 'minimize' function to ensure reasonable structural poses were produced. Any particularly high energy molecules were inspected manually for mistakes in construction and were corrected manually.

As a final step, each database was exported into three commonly used formats: SDF, MOL2, and ASCII (exported as a TXT file), in addition to the MDB format specific to MOE. For internal use, the resulting SMILES strings are stored in a MySQL database allowing for standard Structured Query Language (SQL) searches throughout the collated database internally and as a basis for web-server integration.

Structural Fingerprint Generation. To compare the structural diversity of TIN with the ZINC database, subsets of each database were randomly selected. A pre-prepared ZINC subset of purchasable compounds (#6), that differ from the rest by a Tanimoto cutoff score of 60%, was downloaded. This consisted of 10,082 compounds in SMILES format. 11,000 structures from the 10% diversity subset of TIN were randomly selected to give a representative group of molecules to compare to the ZINC subset. MOE was used to generate initial 3D structures for each molecule. A short java programme was written to calculate the CDK structural fingerprints[14] for each molecule, and sum the occurrence of each fingerprint for plotting.

OpenBabel SMILES searching. The full TIN database can be searched on the 'Search TIN' web page tab using the Java Molecular Editor kindly donated by Peter Ertl (Novartis). Alternatively, a TIN compound can be used as input by typing the compound number (prefixed with RCSI_) in the box provided. Each of these methods inserts a SMILES string in the search input box, which can also be loaded directly with a SMILES sting. After setting the search criteria (Tanimoto similarity score, substructure search, etc), the search starts an OpenBabel command that searches an indexed form of the full TIN database in SDF format (~220 GB file). As this takes a number of hours, the user can logout and will be notified by email when the output files are ready. These are stored in the 'My Results' section with a unique file name created from the 'search description' input.

TIN component parts search. Once the user has selected the TIN compounds they are interested in synthesizing, either through screening the diversity subset or searching the whole dataset using the browser, one can then deconstruct the final synthetic products into their substituent starting materials. These starting materials can all be purchased directly from the supplier, in this case we used Sigma-Aldrich, and the required TIN product synthesized using the chemistry detailed in the methods section and outlined in Schemes 1-12.

On the 'Synthesis' page, one can enter TIN codes (in the form **RCSI_nnnnnnnn**) in the search box provided. The output appears on the page in table format which includes the following columns; 'Sort Order', keycode for the search; 'RCSI Identifier', TIN code number; 'Scaffold Template', core scaffold type; 'Scaffold', full scaffold name; 'Scaffold SMILES', full scaffold structure; 'compound SMILES', reagent structure; 'compound weight', reagent weight; 'Ports', position of port at which R-group is attached; 'RGroups', name of side chain or R-group at specific port; 'RGroup SMILES', side chain/R-group compound structure at specific port; 'RGroups chirality', up to 4 chiral centers were enumerated where the description was not clear; 'RGroup Catalog Number', supplier catalogue number; 'RGroup Supplier', supplier name. Users can then order the compounds directly from the supplier and synthesize the compound using the chemistry outlined and reference in the methods.

RESULTS AND DISCUSSION

TIN aims, as ZINC does, to provide a straightforward way for users to access a large database of novel compounds. Using a 'one-pot' synthetic technique, the combination of unique scaffolds and commercially available compound constituents from the Sigma-Aldrich catalogue, a database of novel synthetic compounds has been constructed. The TIN database is a collection of 28,473,744 virtual compounds, comprising over 220 GB of data in SDF format. Descriptive properties that define drug-like and lead-like compounds are present for all compounds allowing the user to independently search according to these subset properties or to sort databases according to these properties. Of the 28 million compounds, 5,080,762 (17.8%) are considered 'drug-like' (Figures 2, 3, 4) and 1,731,331 (6.1%) are considered 'lead-like'.

To further explore the structural diversity of the TIN database, structural fingerprints were identified in random subsets of both the TIN and ZINC databases. This fingerprint comprises 307 bits, each representing a discrete substructure. A pre-prepared diverse subset of 10,082 ZINC compounds was downloaded, and this was compared to 11,000 compounds chosen at random from the 10% diversity subset of TIN. A table describing the substructure SMARTS pattern of each fingerprint bit, as well as how many molecules from TIN and ZINC contained the substructure is included in Supplementary Table S2.

Out of the 307 fingerprint substructure features, 209 appear in at least one ZINC molecule, and 104 appear at least once in TIN. 102 of these features are common to at least one molecule from ZINC and TIN, while TIN contains two substructures that do not appear in ZINC, and ZINC contains 107 substructures that do not appear in TIN. The distribution of fingerprints is shown in Figure 5. Each point in the figure represents a single substructure feature. It is clear that TIN shares about half the substructure fingerprints with ZINC, with a broadly similar distribution. Fingerprints lying along the x-axis are those fingerprints in ZINC which do not occur in TIN. These comprise about half of the total fingerprint substructures contained in ZINC. Referring to Supplementary Table S2, the bulk of the features in ZINC that do not exist in TIN relate to various phosphorus, sulfur, and halogen containing

groups. However, phosphorus and halogen containing groups are reactive functional groups which may not be desirable from a medicinal chemistry point of view.

The large size of the TIN database restricts online downloading of the full dataset in 3D format, but all 28 million molecules are freely available for download in isomeric SMILES format. A smaller diversity subset (10%) was created and is available for download in SDF, MOL2, MDB and SMILES formats. 'Lead-like' and 'drug-like' subsets are also freely available for download in SDF, MOL2, MDB and SMILES formats. Once these have been used in virtual screening studies, it is envisaged that researchers will synthesize the hit compounds, and test them *in vitro* for effectivity.

The TIN website also has a web-based application which allows users to search for molecules similar to the input molecule. Once lead small molecule inhibitors are established, the full TIN database can be searched for readily synthesizable derivatives using the JME molecule editor and the OpenBabel search function. These derivatives can then also be synthesized and screened, potentially yielding novel therapeutically useful drugs.

TIN is a user-friendly public access database that provides a comprehensive search of small molecules available through multicomponent synthesis methods. Users are able to search the TIN database based on a variety of criteria; (i) diversity subset databases that contain a representative 10% of each scaffold database are available for ready download, (ii) drug-like compounds that obey Lipinski's 'Rule-of-Five', (iii) smaller lead-like compounds, or (iv) all compounds (although this is only available in SMILES format due to the data transfer limitations. The issue of reducing the size of chemical space by excluding possible solutions is addressed by allowing TIN to be fully searchable; the structure based searches performed by users allow for full exploration of the entire collection of molecules and understanding of all TIN's molecular structures. Thus if a 'hit' is found using the 10% diversity subset, the remaining 'near neighbors' (that were filtered out) can now be accessed using the structure-based search engine. Each molecule is available for download in a variety of formats allowing for use in a variety of molecular modelling programs.

TIN is a sizable database containing a high number of unique molecules which satisfy 'drug-like' and 'lead-like' properties all of which are theoretically synthesizable. The open access nature of this large resource will hopefully facilitate further research from a variety of sectors in virtual screening and future drug design and development.

To investigate how successfully the building blocks of the TIN database have been shown to act against protein targets *in vitro*, a substructure search for each of the 12 substructure scaffolds was carried out on the open-access ChEMBL[20] and PubChem[21] databases, which contain data on the biological activity of small molecules. The results are summarized in Table 2. This shows that a total of 1,403 molecules containing TIN substructures have been found, with 37 experimentally validated activities against 18 protein targets (not including possible duplicates between databases). This indicates that there is potential for finding bioactive compounds within TIN. However, considering that TIN itself consists of over 28 million compounds, finding only 1,403 tested compounds in publically available databases indicates that TIN represents a relatively unexplored region of the chemical space.

Recent analysis of modern patenting trends by the four largest pharmaceutical companies has shown a trend towards increasing hydrophobicity,[17] and this is likely to result in higher attrition in late-stage clinical testing. Similarly, GlaxoSmithKline have recently made available the results of their anti-malarial high throughput screening assays, in which they found a median molecular weight of 446 Da for active compounds.[22] By the nature of the conjugated styrylisoxazole ring centre scaffold, the TIN database contains a significant proportion of highly hydrophobic molecules which fail the Lipinski Rule-of-Five criteria (Figure 2). This is clearly inconsistent with the aim of the TIN database, namely the identification and development of novel pharmaceutical drugs. However, the widened scope of compounds considered a lead worthy of patent protection by 'big pharma',[17] coupled with large average size of anti-malarial hits found by GSK, implies that the TIN database should still hold valuable starting leads and novel scaffolds. Furthermore, the versatility of the multicomponent synthesis allows one to readily improve polarity and drug-likeness by replacing components with more polar analogues. A simple WWW-based structure search of the lead compound will output all possible analogues in TIN,

which can easily be sorted for non-violation of Lipinski's rules of five (where 'lip_druglike' = 1), thus returning the user to more useful, low lipophilicity, chemical space.[17]

To illustrate the potential utility of the TIN database, we searched it for compounds structurally similar to the anti-thrombotic platelet integrin inhibitor **Tirofiban** (Figure 6a). The top ranked result was the compound **RCSI_00000319** (Figure 6b). The overlapping conformations were calculated using MOE (CCG), and show surprisingly good alignment with the key pharmacophore features in **Tirofiban** (Figure 6c). The compound '**319**' is part of the **MA01** scaffold group (Scheme 1), with the diketone (**K**) option **K01** and pyrazole **X02** option used to build the azole ring, while the aldehyde used contributes major part of the molecule, the phenoxy-propane-dimethylamine. Further derivatization is thus possible at each of these positions, offering the possibility of readily synthesizing novel anti-thrombotics. The authors envisage that the TIN database and server will find similar utility in converting substrates into novel drug leads.

CONCLUSIONS

We have produced a virtual combinatorial library of 28 million molecules that are all theoretically synthetically feasible. The chemistry involved in synthesizing each of the keywords is well established and previously published. The databases are freely available and can be downloaded either (i) in their entirety, (ii) as 10% diversity subsets, (iii) as drug-like molecule subsets obeying Lipinski's Rule-of-Five, (iv) as lead-like molecule subsets, or (v) as subsets defined by the core scaffold molecule.

Furthermore, we have created an online search interface that allows users to draw input molecules for substructure or similarity searching of the full database. The 2-dimensional search function can also be accessed using 'TIN' codes of molecules that exist in the database. Thus users can find structural analogues of hit molecules found using TIN, such as the diversity or drug-like subsets.

This facility will allow users to rapidly convert lead molecules into synthetically feasible compounds for synthesis and *in vitro* testing. Once hit compounds are identified, the variability inherent in the multicomponent synthesis methods will allow for relatively rapid synthesis of a wide range of

derivatives, which can then be used in repeating cycles of drug discovery. Further scaffold hopping is available using 2D searching (substructure/similarity) of the full database. The TIN database can be used in conjunction with the extremely versatile ZINC database of commercially available compounds.[7] Lead molecules found using virtual screening of the ZINC compound library that are purchased and tested *in vitro*, can then be converted into synthetically feasible analogues using the 2D search function in TIN. This will then give researchers access to chemical diversity beyond that of the finite commercially available compound library.

ACKNOWLEDGEMENT

We thank Andrew Henry, Niall English, Guido Kirsten and Emilio Esposito of Chemical Computing Group, Montreal for MOE programming assistance. We thank Peter Ertl of Novartis for the generous donation of the Java Molecular Editor applet. We thank Denis Shields, and the reviewers, for useful comments regarding the manuscript. AJC was funded by the Irish Higher Education Authority 'Programme for Research in Third-Level Institutions, Cycle 3' and Irish Research Council for Science, Engineering and Technology. KVD was funded by the Irish Health Research Board. FJD was funded by Science Foundation Ireland.

Supporting Information Available.

Table S1: MOE Descriptors calculated for each TIN molecule.

Table S2: Structural Fingerprints calculated for TIN and ZINC subsets.

Table S3: Bioactive molecules using TIN scaffold substructures.

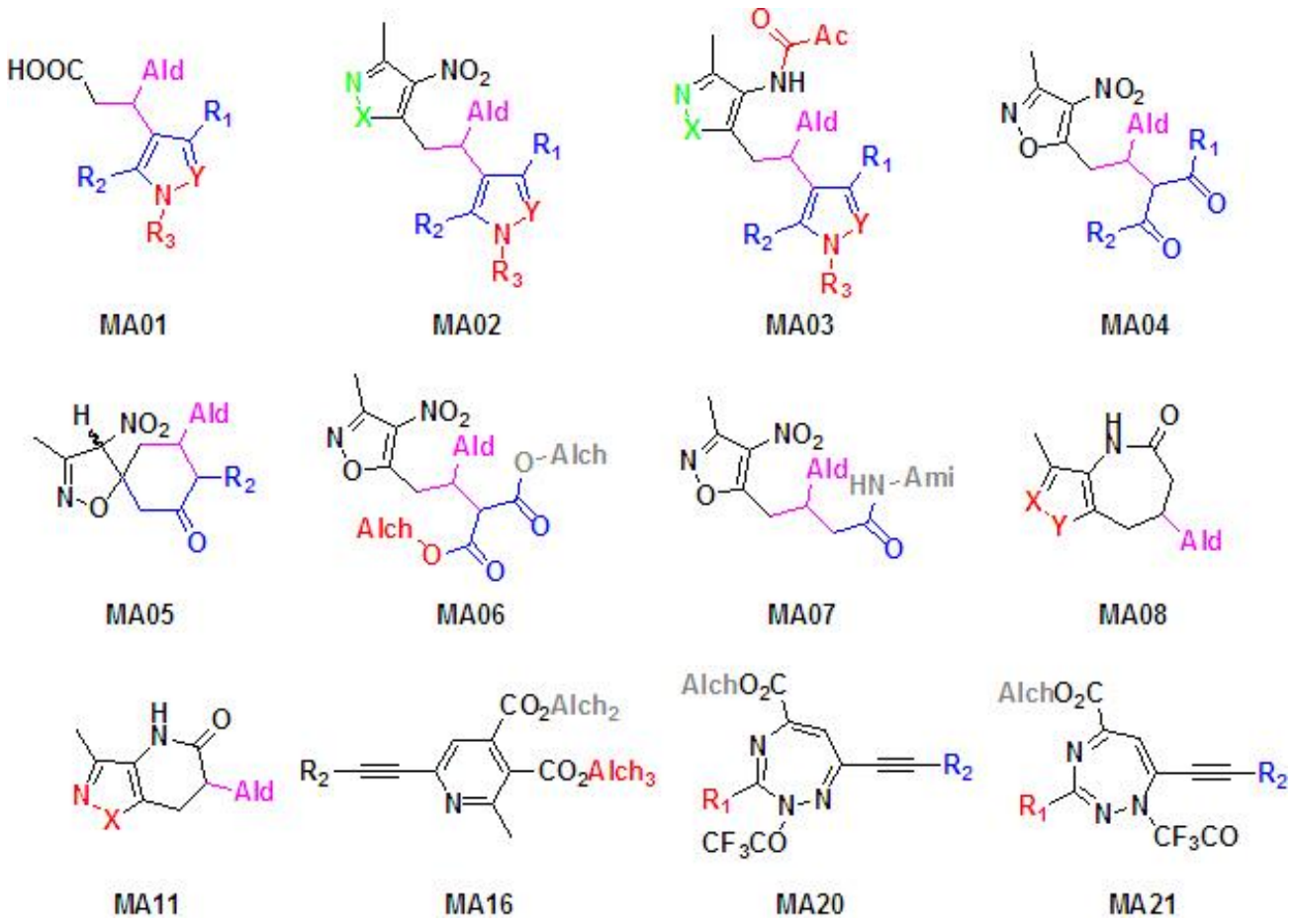


Figure 1: Summary of scaffolds enumerated in the TIN database.

Substituent colors are explained in **Schemes 1-12**.

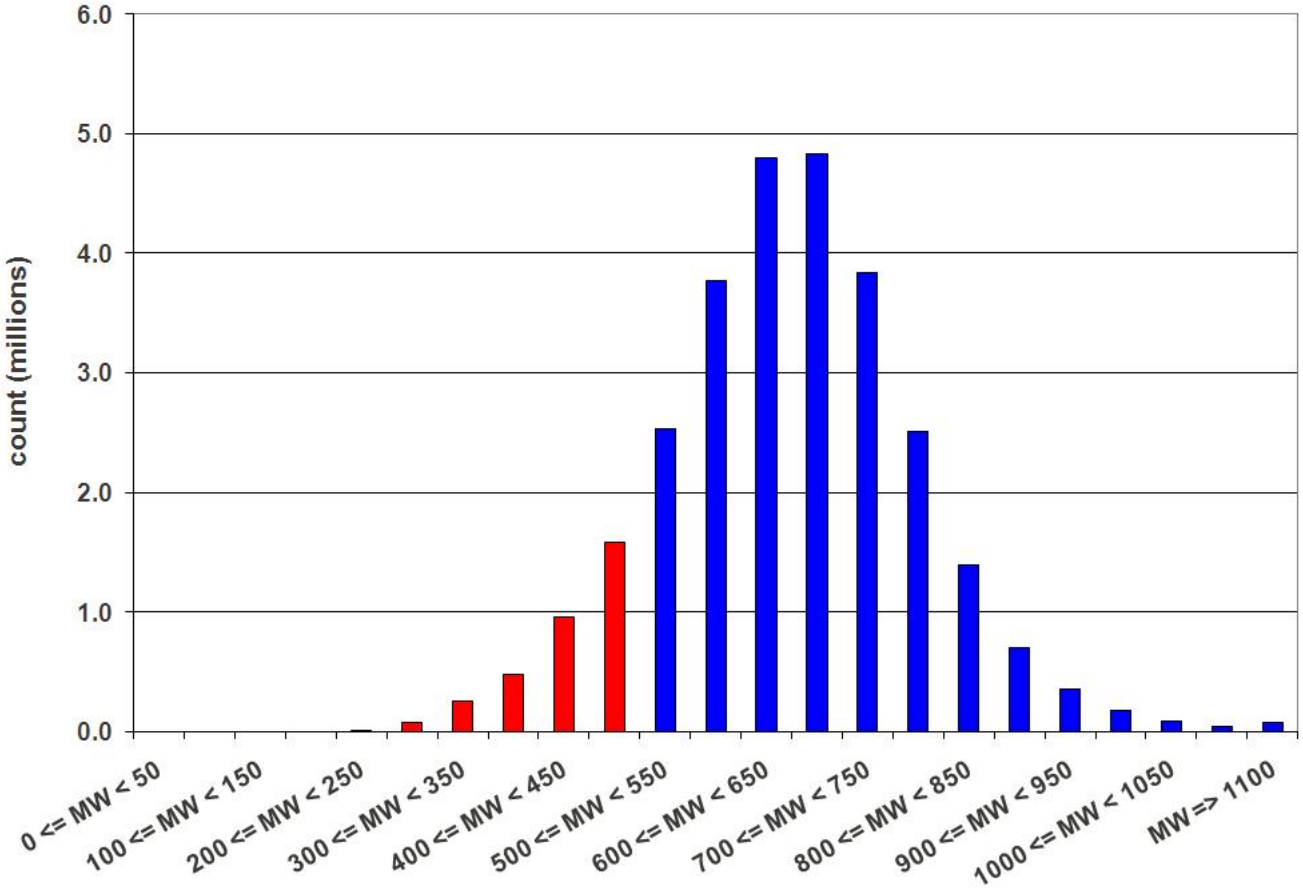


Figure 2: Histogram of TIN compounds by molecular weight.

'Drug-like' compounds are depicted with red bars.

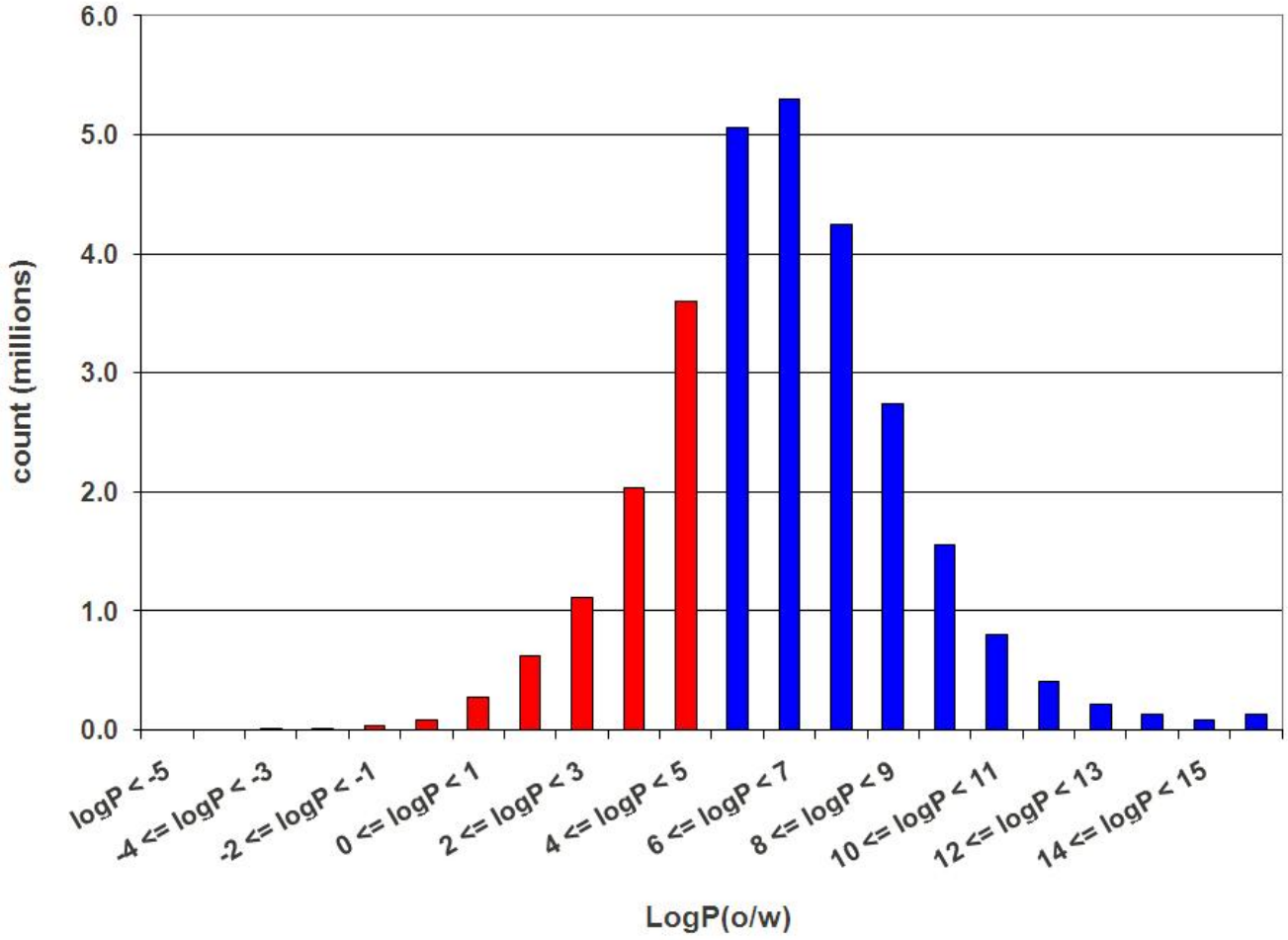


Figure 3: Histogram of TIN compounds by logP(o/w).

'Drug-like' compounds are depicted with red bars.

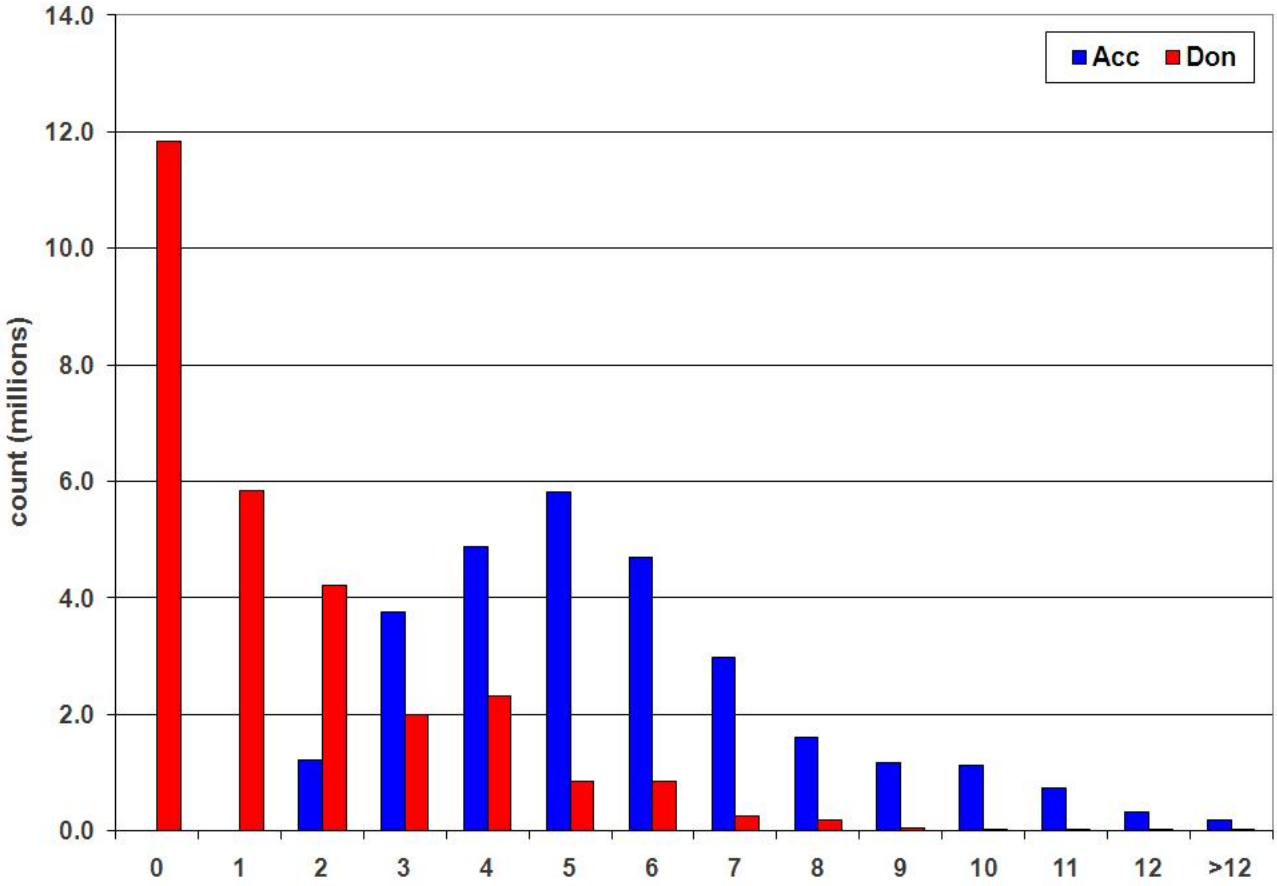


Figure 4: Histogram of hydrogen-bond acceptors and donors.

Hydrogen-bond acceptors are depicted with blue bars, while Hydrogen-bond donors are depicted with red bars.

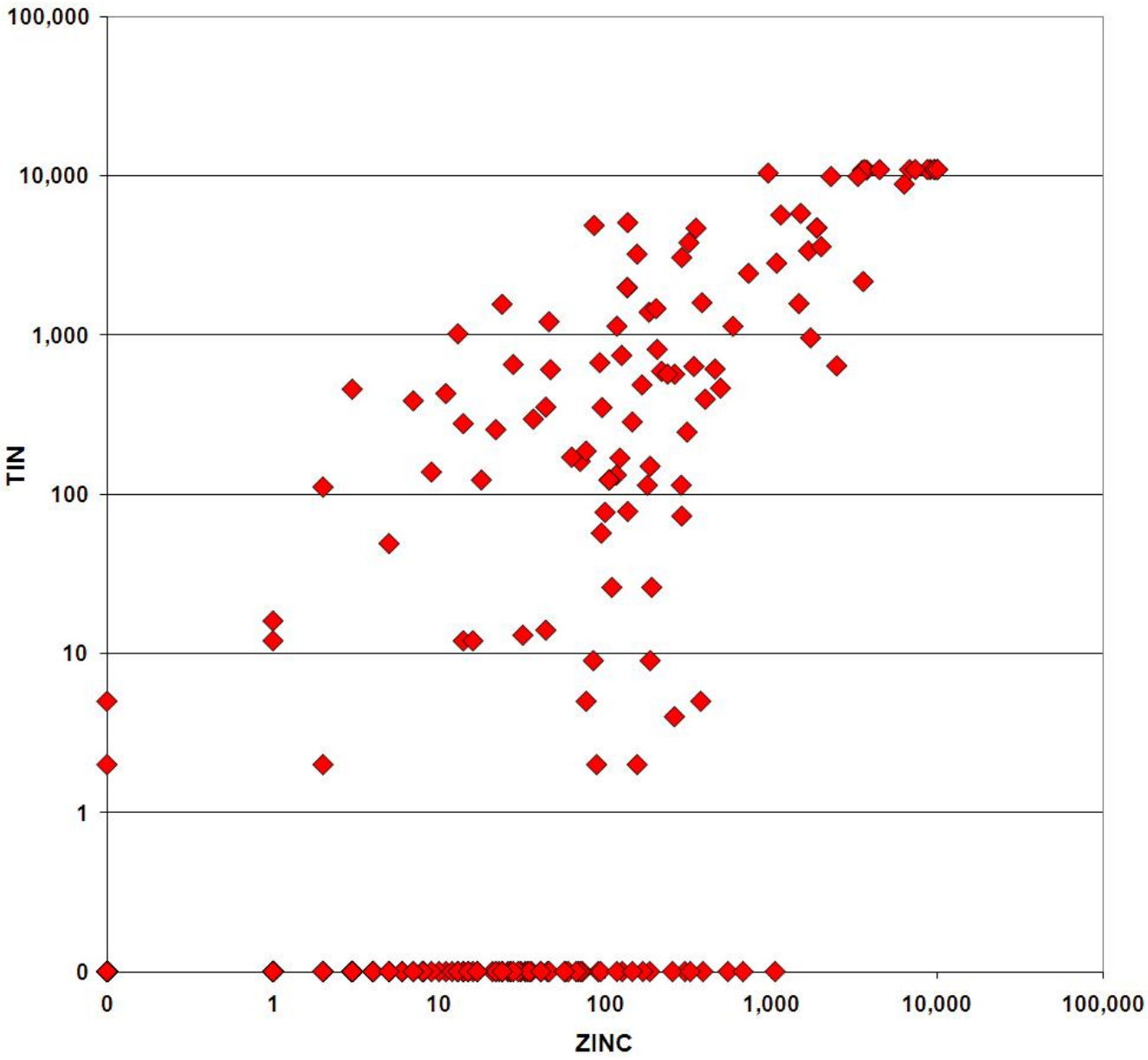


Figure 5: Diversity of TIN compared to ZINC.

11,000 compounds were randomly selected from the diversity subset of TIN and the Tanimoto 0.6 subset of ZINC (purchasable compounds). Substructure fingerprints were analyzed in each compound using a 307 bit search. Each point in this logarithmic scatter plot thus corresponds to the number of compounds found to have at least one copy of the fingerprint in either the ZINC dataset (x-axis) or TIN dataset (y-axis). To allow graphical representation, "null" is represented by "0.1".

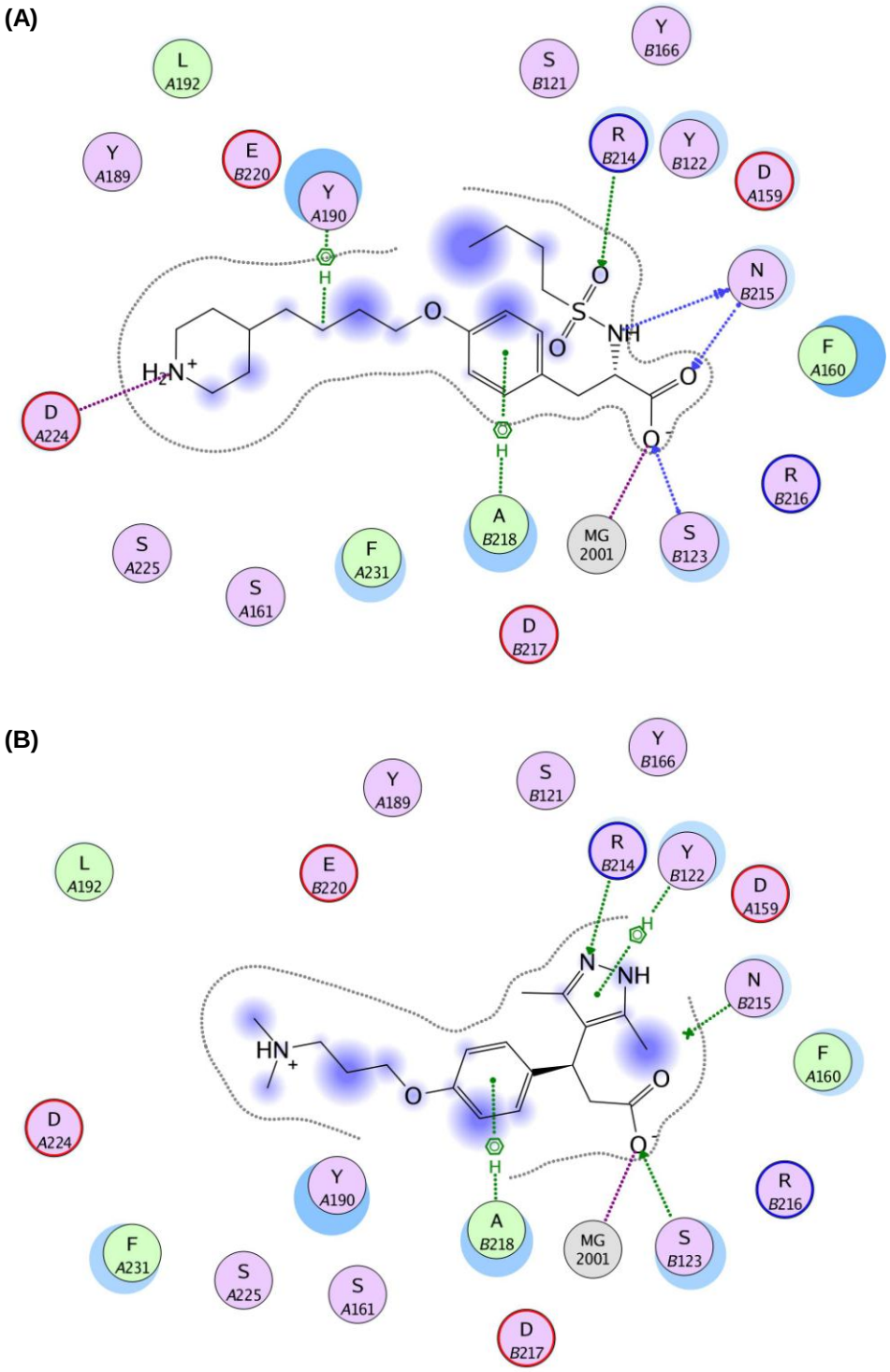
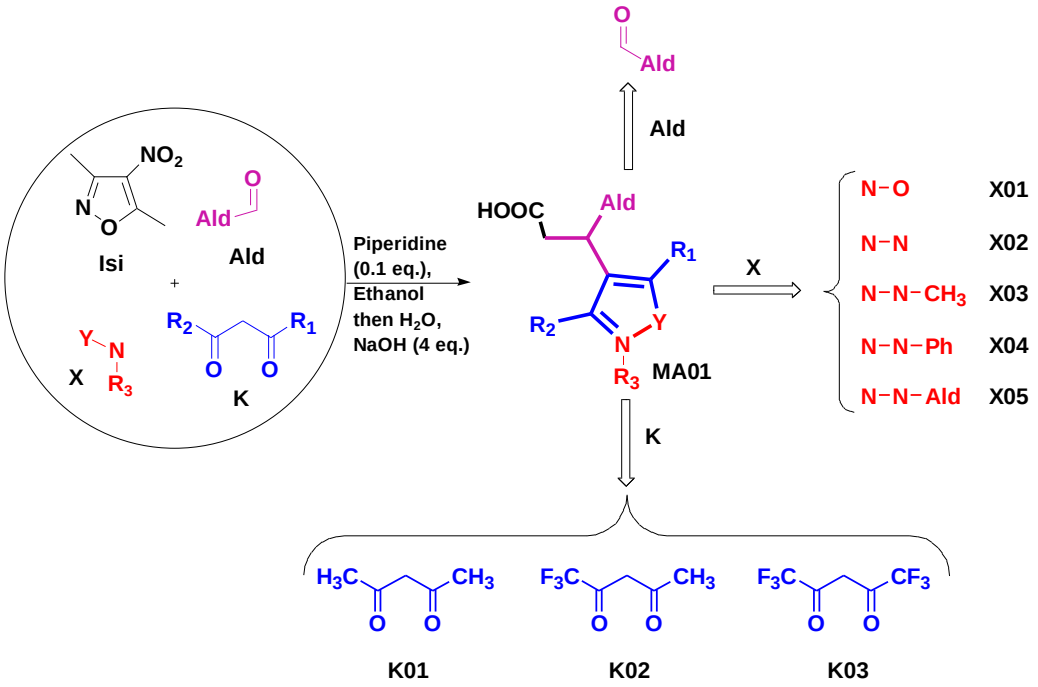
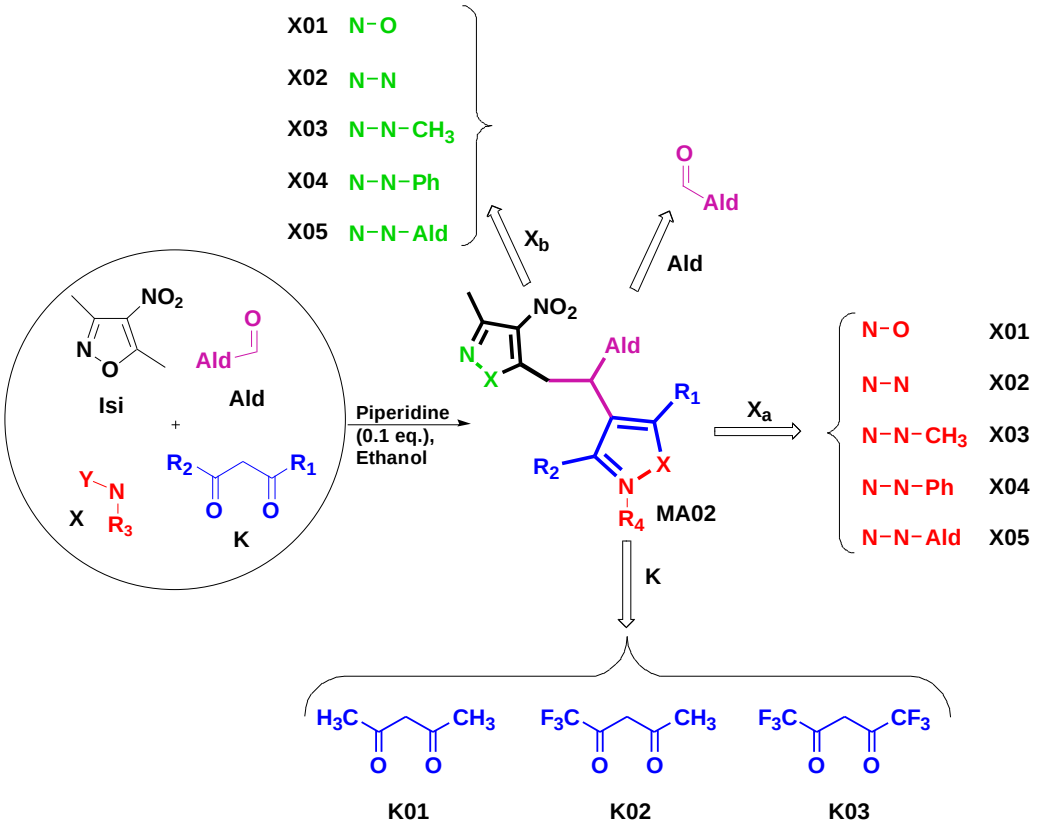


Figure 6: TIN search using Tirofiban.

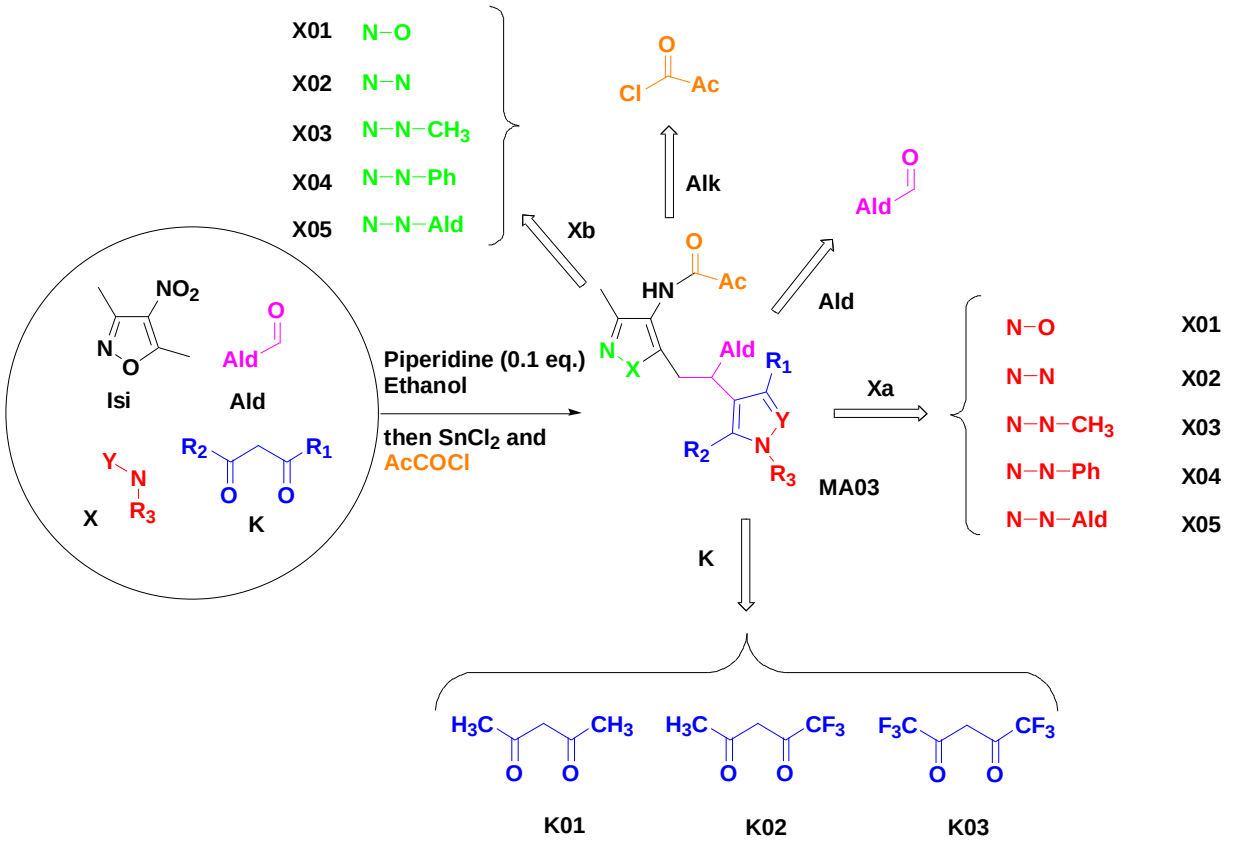
A schematic of the structure of Tirofiban bound to platelet integrin GPIIbIIIa (2VDM) is shown in (A), while that of the TIN hit result '319' is shown in (B). This shows that both ligands interact with residues Arg214, Ser123, Ala218 and the Mg2+ cation. The MOE ligand interactions color scheme is as follows: polar residues, pink; acidic, red ring; basic, blue ring; hydrophobic residues, green; solvent exposure, blue zone; H-bonds, dashed green lines.



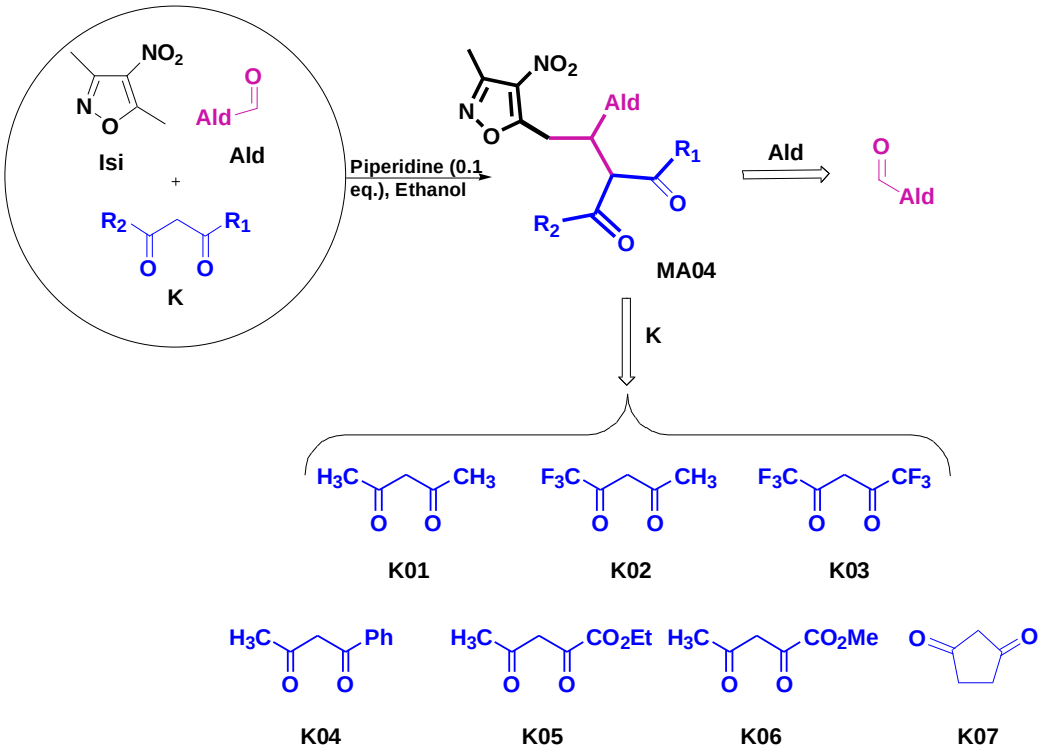
Scheme 1: Synthesis and definition of scaffold MA01.



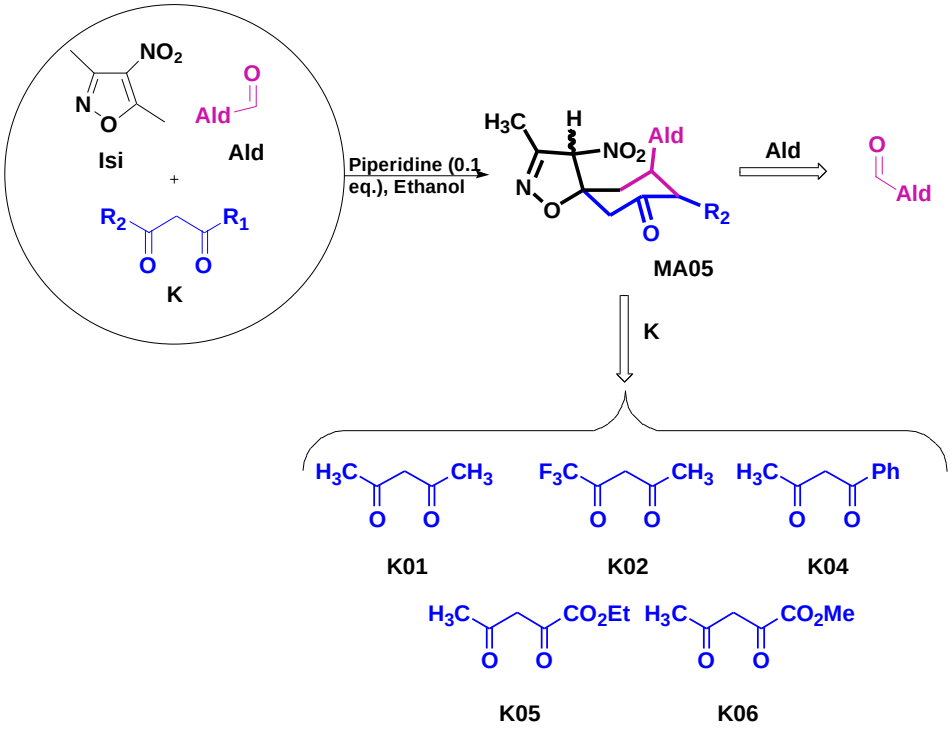
Scheme 2: Synthesis and definition of scaffold MA02.



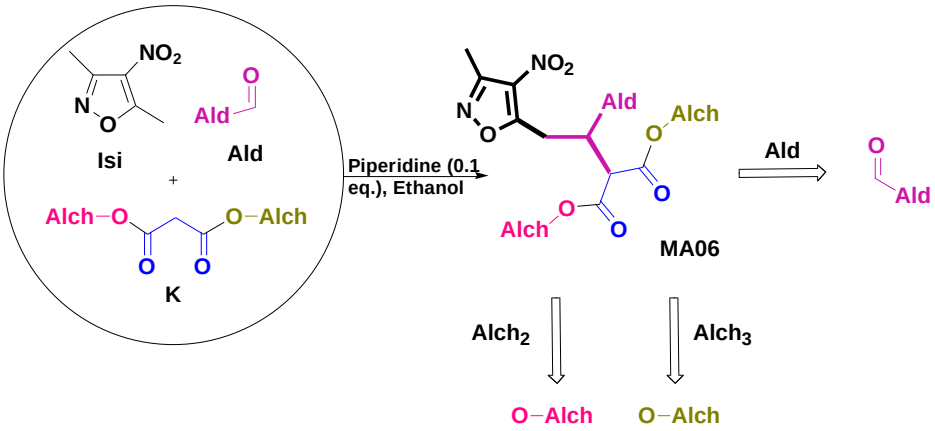
Scheme 3: Synthesis and definition of scaffold MA03.



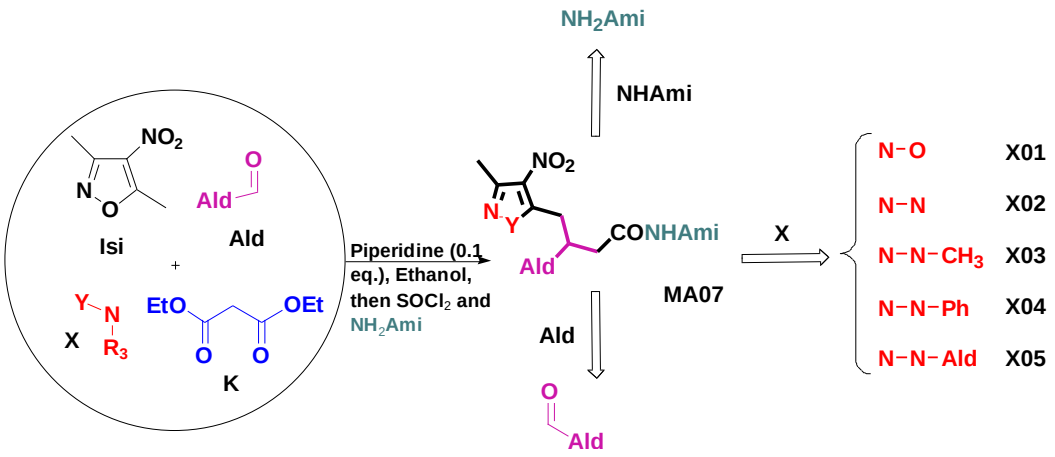
Scheme 4: Synthesis and definition of scaffold MA04.



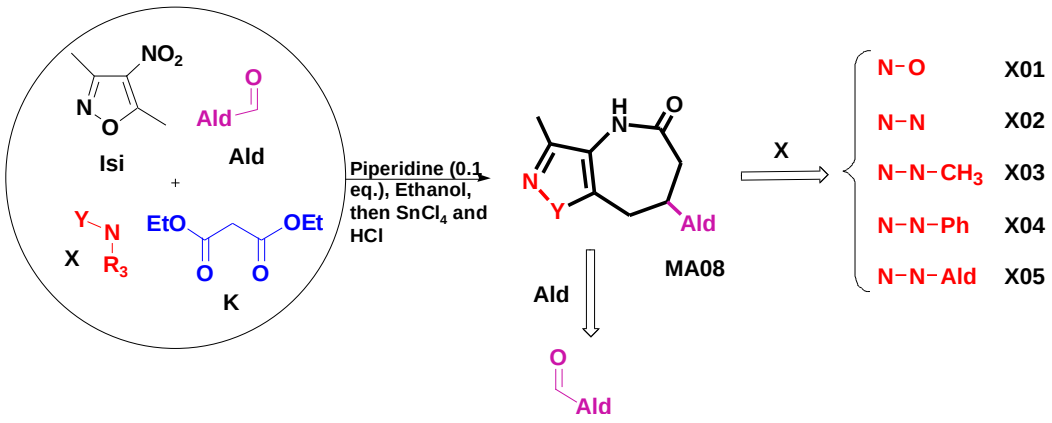
Scheme 5: Synthesis and definition of scaffold MA05.



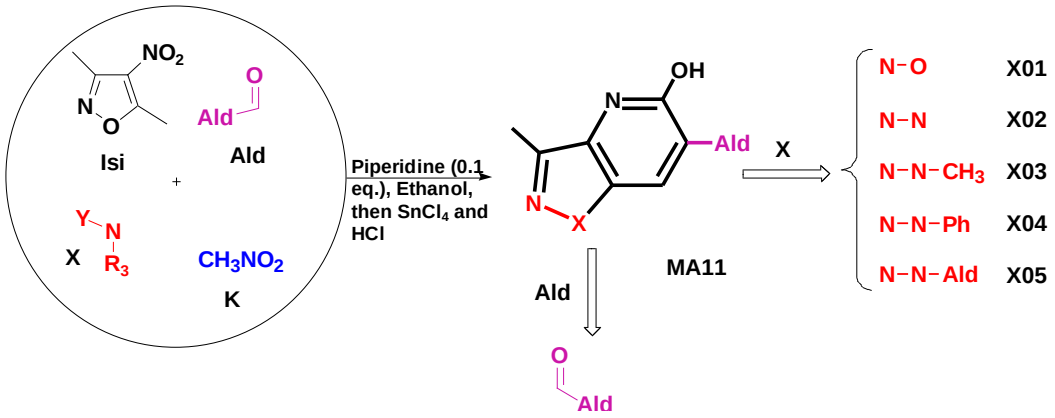
Scheme 6: Synthesis and definition of scaffold MA06.



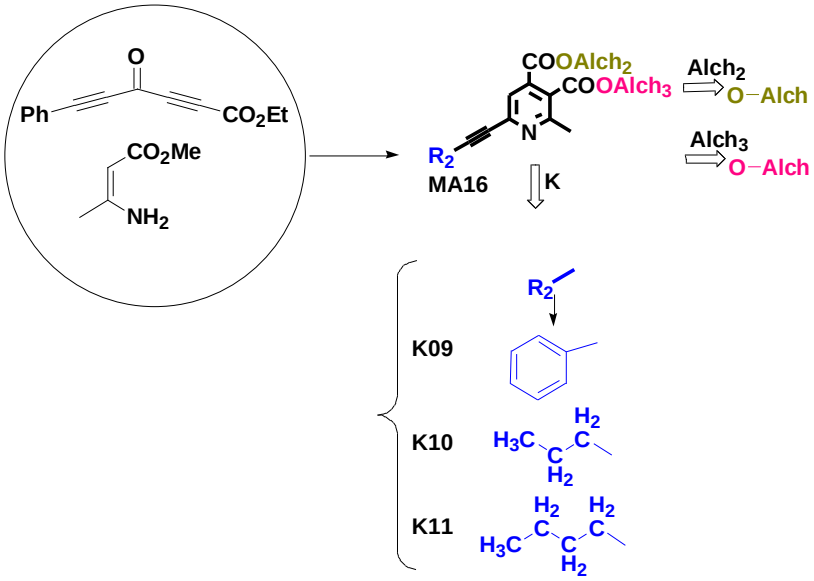
Scheme 7: Synthesis and definition of scaffold MA07.



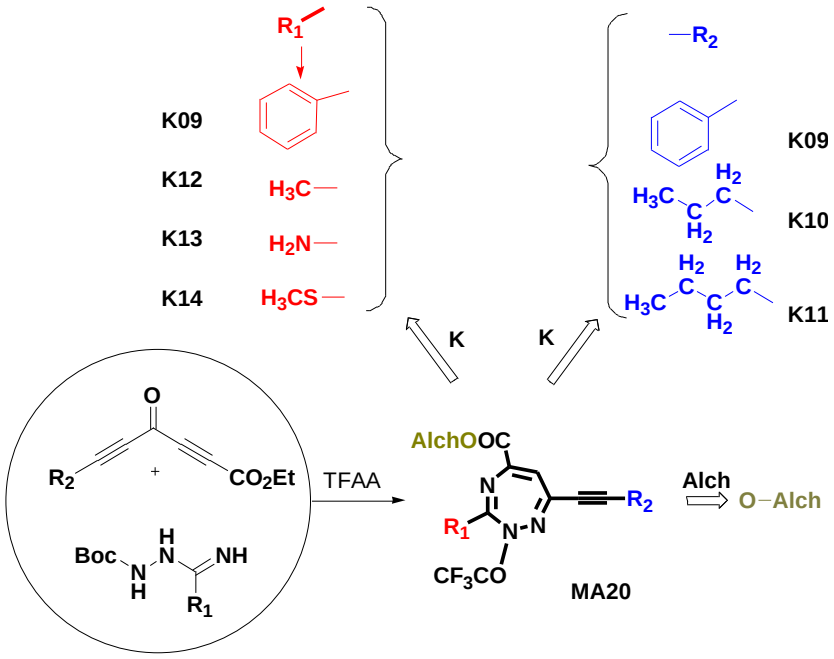
Scheme 8: Synthesis and definition of scaffold MA08.



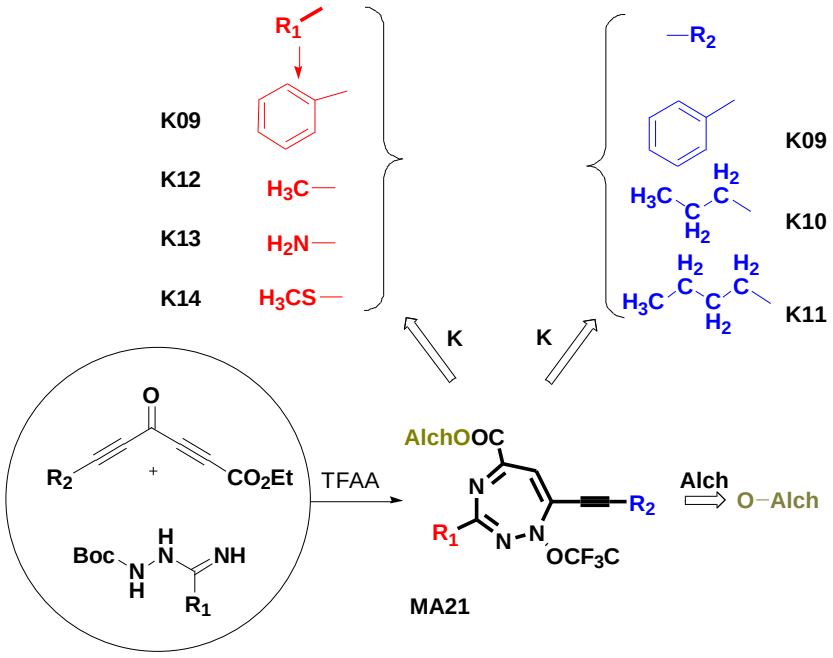
Scheme 9: Synthesis and definition of scaffold MA11.



Scheme 10: Synthesis and definition of scaffold MA16.



Scheme 11: Synthesis and definition of scaffold MA20.



Scheme 12: Synthesis and definition of scaffold MA21.

TABLES.

Table 1: Reactant component subsets.

Type of R-group†	Abbreviation	Reactive group SMARTS‡	Leaving Group	Leaving Group SMARTSΨ	Number of Entries
Aromatic Aldehyde	Ald	[CH](=O)c	Aldehyde	[CH](=O)	226
Aromatic Alcohol	Alch_aro	[OH]C([!O])c	Hydroxyl	[OH]	120
Aromatic Alcohol	Alch_aroH	[OH]C([!O])c	Hydrogen	[H]	120
Aliphatic Alcohol	Alch_naro	[OH]C([!O])C	Hydroxyl	[OH]	404
Aliphatic Alcohol	Alch_naroH	[OH]C([!O])C	Hydrogen	[H]	404
Secondary Aromatic Amine	NHAmi_aroNH	[NH]c	Hydrogen	[H]	112
Primary Aromatic Amine	NHAmi_aroNH2	[NH2]c	Hydrogen	[H]	386
Secondary Aliphatic Amine	NHAmi_naroNH	[NH]C	Hydrogen	[H]	97
Primary Aliphatic Amine	NHAmi_naroNH2	[NH2]C	Hydrogen	[H]	142
Aromatic Carboxylic Acid	Ac_aro	[CX3](=O)(O) c	Carboxylic Acid	C(=O)O	616
Aliphatic Carboxylic Acid	Ac_naro	[CX3](=O) (O)C	Carboxylic Acid	C(=O)O	324

†Selected molecules in the Sigma-Aldrich catalogue were classified into possible reactant subunits compatible with multicomponent chemistry.

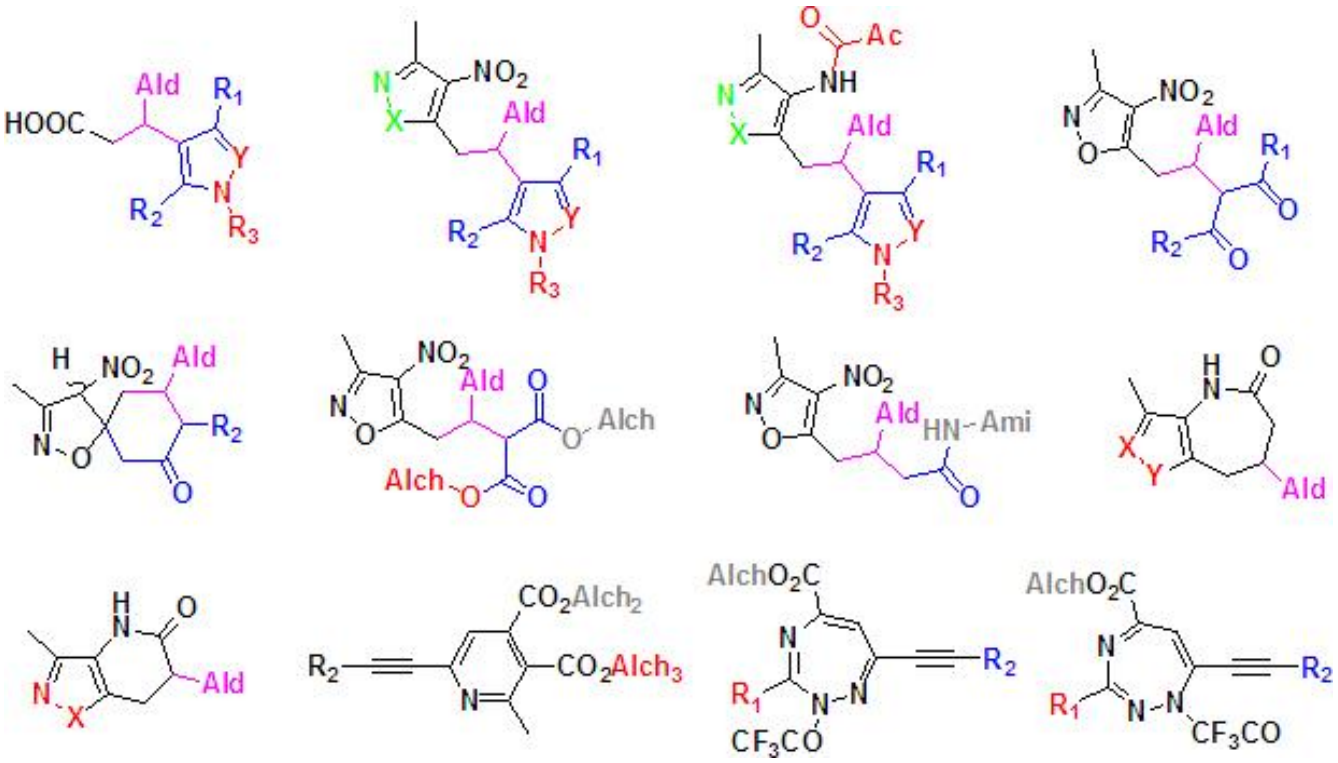
‡'Reactive Groups' were identified, but only a subset of this was removed (the 'Leaving Group'Ψ) in preparation for linking to scaffolds.

Table 2: Bioactive molecules with TIN scaffolds in the ChEMBL and PubChem databases.

	TIN scaffold substructure	**Active compounds***	**Protein Targets**
ChEMBL	35	26	17
PubChem	1368	11	1
Total	1403	37	18

*PubChem active compounds are defined as those which pass the PubChem filter "Active in any bioassay". ChEMBL active compounds are those that show a result in any assay that is Activation/Inhibition (IC50, EC50, Ki) < 10000 nM or Max Inhibition, Activity, Recovery > 50%.

SYNOPSIS TOC:



TIN is a virtual combinatorial database enumeration of diversity-orientated multicomponent syntheses, containing over 28 million compound structures for searching or downloading.

REFERENCES AND NOTES:

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004,** *303*.
(2) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004,** *432*, 862-865.
(3) Tse, M. Lead identification: Combining strengths to find high-quality leads. *Nat. Rev. Drug Discov.* **2010,** *9*, 593
(4) Ferreira, R.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B.; Keiser, M.; McKerrow, J.; Maloney, D.; Irwin, J.; Shoichet, B. Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. *J. Med. Chem.* **2010,** *53*, 4891-4905.
(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004,** *3*, 935-949.
(6) Oprea, T.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004,** *8*, 349–358.
(7) Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005,** *45*, 177-182.
(8) Adamo, M. F. A.; Donati, D.; Duffy, E. F.; Sarti-Fantoni, P. Multicomponent synthesis of spiroisoxazolines. *J. Org. Chem.* **2005,** *70*, 8395-8399.
(9) Adamo, M. F. A.; Adlington, R. M.; Baldwin, J. E.; Pritchard, G. J.; Rathmell, R. E. Practical routes to diacetylenic ketones and their application for the preparation of alkynyl substituted pyridines, pyrimidines and pyrazoles. . *Tetrahedron* **2003,** *59*, 2197-2205.
(10) Adamo, M. F. A.; Baldwin, J. E.; Adlington, R. M. Application of bis-acetylenic ketones in synthesis: one-pot preparation of the 1,2,4-triazepine and oxatriazaindenone cores. . *J. Org. Chem.* **2005,** *70*, 3307-3308
(11) Adamo, M. F. A.; Donati, D.; Duffy, E. F.; Sarti-Fantoni, P. Modular synthesis of isoxazoleazepinones and pyrazoleazepinones *Tetrahedron* **2007,** *63*, 2684-2688.
(12) Adamo, M. F. A.; Duffy, E. F. Multicomponent synthesis of 3-heteroarylpropionic acids *Org. Lett.* **2006,** *8*, 5157-5159.
(13) Adamo, M. F. A.; Duffy, E. F.; Donati, D.; Sarti-Fantoni, P. Modular synthesis of isoxazolopyridones and pyrazolopyridones. *Tetrahedron* **2007,** *63*, 2047-2052.
(14) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des* **2006,** *12*, 2111-2120.
(15) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Computer Methodology* **1990,** *3*, 537-547.
(16) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* **2003,** *21*, 449-462.
(17) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **2007,** *6*, 881-890.
(18) Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997,** *23*, 3-25.
(19) Oprea, T. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000,** *14*, 251–264.
(20) Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J Comput Aided Mol Des* **2009,** *23*, 195-198.
(21) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An overview of the PubChem BioAssay resource. *Nucleic Acids Res* **2010,** *38*, D255-266.

(22) Gamo, F. J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J. L.; Vanderwall, D. E.; Green, D. V.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010,** *465*, 305-310.